DAO Office Note 96-15

Office Note Series on Global Modeling and Data Assimilation

Richard B. Rood, Head
Data Assimilation Office
Goddard Space Flight Center
Greenbelt, Maryland

GEOS/DAS Quality Control Strategy Document Version 1.1

Dick P. Dee[†]
Alice R. Trenholme[†]

Data Assimilation Office, Goddard Laboratory for Atmospheres
[†] *General Sciences Corporation, Laurel, Maryland*

Goddard Space Flight Center
Greenbelt, Maryland 20771
August 1996

Abstract

This document outlines the basic strategy for developing and implementing GEOS/DAS QC, the future quality control component of the data assimilation system currently in development at the DAO. Our primary purpose is to define a conceptual basis for design and to provide rough guidelines for implementation, with a view toward long-term objectives. The first steps that must be taken in order to get an operational, self-contained QC component into GEOS/DAS 2.1, and subsequent developments required for GEOS/DAS 3, are described in some detail. This document will be followed by a requirements document containing a functional definition of the components of the GEOS/DAS QC.

This is Verson 1.1 of this document which supersedes Version 1.0 (June 1996). The only change to the previous version is that Figures 2 and 3 have been redrawn and slightly revised.

Contents

List of Figures

1 Introduction

This document outlines the basic strategy for developing and implementing GEOS/DAS QC, the future quality control component of the data assimilation system currently in development at the DAO. The strategy is driven by the following two fundamental long-term requirements:

- *GEOS/DAS QC must be state-of-the-art, both from a scientific and from a software-engineering point of view.*
  - The DAO analysis system is premised on the ability to characterize observation errors and to quantify them statistically, which obviously cannot be done well unless gross errors can be identified and effectively removed. (The converse is true as well).
  - Operational experience appears to indicate that QC has as much impact on analysis accuracy as any other major component of an assimilation system.

- *GEOS/DAS QC must be self-contained and logistically independent of QC systems in place at other operational centers.*
  - DAO cannot afford to be subject to operational decisions at other centers which are beyond our control.
  - DAO needs to have intellectual control over the QC system (both in terms of algorithm and software design and code) in order to be able to incorporate QC components for new data types.

Before entering into particulars, we briefly describe the quality control process in general terms. The concepts and general approach are strongly influenced by the quality control system that has evolved at the National Centers for Environmental Prediction as documented in Ballish, 1991, Collins, 1991, Collins and Gandin, 1996, 1995, 1992, 1990, Gandin, 1991, Kalnay *et al.*, 1996, Morone, 1991, and Woolen, 1991.

2 Quality control

Throughout this document, *quality control* (QC) refers to the process by which observational data and their attributes are analyzed in order to

- identify data items which are likely to contain *gross errors*;
- attempt to correct such errors.

Gross errors (also known as *rough errors* in papers by Collins and Gandin, 1990-1996) are any inaccuracies that cannot be explained in terms of

(1) detection noise of the observing instrument under normal operating conditions,
(2) small-scale variability of the atmosphere, and
(3) approximations inherent to the observation operator.

Clearly QC for any single observation must involve information other than the observational datum itself, such as:

- attributes of the observation (units,time, location, etc.)
- observation error statistics (for non-gross errors)
- operating conditions of the instrument (e.g. scan angle)
- gross error statistics
- probable causes of gross errors
- other nearby observations

- known spatial and/or temporal relationships between atmospheric variables
- climatological information
- a model forecast valid in the vicinity of the observation location
- forecast error statistics

The QC process consists of a set of algorithms which examine each data item, singly or jointly, in the context of this additional information. Their primary purpose is to determine which of the data are likely to contain unknown (incorrigible) gross errors, and which are not. The algorithms can be categorized as follows:

1. A *testing algorithm* produces a *quality mark* for each data item it processes. Each test can be regarded as a hypothesis test on the actual error associated with the data item in view of some of the additional information listed above.

2. A *correction algorithm* produces an estimate of the actual gross error, which may (or may not) be subsequently used to modify a data item. Correction algorithms analyze the probable cause of the gross error, and produce an estimate of the gross error if the cause can be determined with a sufficient degree of confidence.[1]

3. A *decision making algorithm* (DMA) produces the final decision regarding the disposition of each data item, which is one of the following: `accept`, or `reject`. This decision is made at the very end of the QC process, based on the cumulative results of the testing and correction algorithms (TCA) to which the data have been subjected.

[1] It might seem natural to include *bias correction algorithms* in the domain of quality control as well. However, bias is not a manifestation of gross error: bias is, by definition, systematic whereas gross errors tend to be erratic. Conceptually, therefore, bias correction must be separated from the quality control process.

3 Modular structure of the QC process

Figure 1 depicts the boundaries and the basic modularity of the future GEOS/DAS QC. The information flow is from top to bottom.

DAO will receive observational data, possibly supplemented with QC marks, from various suppliers at different time intervals. See DAO Office Note 96-16 *(Data Assimilation Computing and Mass Storage Requirements for 1998)* for a detailed description of the expected operational data streams. Synchronization and reformatting of the data takes place prior to ingestion by the QC system. At this point, the data and their generic attributes will be contained in ODS (Observational Data Stream) files. Data-type specific attributes will be contained in OMS (Observational Meta-Data Stream) files. See DAO Office Note 95-01 *(Documentation of the GEOS/DAS Observation Data Stream (ODS) Version 1.0)* for a definition of these internal DAO data formats. These data formats will be used throughout the GEOS/DAS system to store observational data and their attributes, including quality marks. In particular, *the QC component of the system relies primarily on ODS,* both for input of observational data and its attributes, and for output of quality marks and QC decisions. Monitoring the output of the QC process is one of the main functions of DAO's On-Line Monitoring System (DOLMS); see DAO Office Note 96-13 *(Requirements for DAO's On-Line Monitoring System (DOLMS) Version 1.00).*

We reiterate the *scope of the QC process* as defined in general terms in section 2. The QC process does *not* include data synchronization and reformatting, nor does it include actual reporting functions other than simply tallying and administrating

QC marks and decisions. Conceptually, QC also does not include *bias correction* or *super-obbing* (combining several observations into a single 'super-observation'), since its function is strictly limited to the identification and estimation of gross errors. Notwithstanding this conceptual boundary, the design of bias correction and super-obbing algorithms must be closely related to the design of QC modules.

The GEOS/DAS QC modules consist of *testing and correction algorithms* (TCA) and *decision making algorithms* (DMA). The DMA takes place just prior to analysis, based on all available information from the TCA. Within the TCA, a conceptual distinction is made between algorithms which require a forecast valid for the time of observation *(post-forecast TCA)*, and those which do not *(pre-forecast TCA)*. From an operational point of view, a distinction must be made between *on-line TCA* (which will be embedded within the analysis component of an assimilation system) and *off-line TCA* (which will not be so embedded). The conceptual and operational distinctions are not necessarily the same, however, and in fact they should be flexible in order to support different modes of operation (see section 6 below). If post-forecast TCA utilizes, say, a 6-hour forecast based on the latest analysis computed by the system, then it must be implemented on-line. Alternatively, post-forecast TCA could be supplied with forecast information otherwise obtained, for example by a previous run of the system. In that case post-forecast TCA may be implemented off-line. Generally, the QC system should be designed to be able to accept forecast information from off-line sources for post-forecast TCA.

The vertical cuts through the TCA components in Figure 1 indicate the dependence of many of the QC algorithms on data type. Particularly at the upstream end of the information flow, QC mostly relies on information specific to a particular data type.

Moving downstream in the QC process, tests tend to become increasingly generic, as information from different data sources is combined and analyzed jointly. An example of a generic test is a *buddy check*, which can act as the final downstream QC filter. We will discuss the buddy check in some detail below.

4 Transition strategy

The QC system described in the previous section represents our long-term objective. QC software currently in place at DAO is rudimentary at best; the current version of the data assimilation system (GEOS/DAS 2.0) strongly relies on QC performed at the National Centers for Environmental Prediction (NCEP). Further on in this section we will refer to the next two major releases of GEOS/DAS: versions 2.1 and 3.0. Specifications for these releases outside the sphere of quality control will be described elsewhere.

Regarding quality control, our strategy for moving from the current situation to the long-term objective is based on the following (short-term) priorities:

- *A self-contained QC system should be operational at DAO as soon as possible.*
  - Until this is the case, we will not be assured of system stability, since the QC system at NCEP is in a state of flux.
  - We need to get hands-on experience before we can actually begin to improve QC.
  - We need the ability for new data types to flow through the system as soon as possible, even if they are not (yet) being assimilated *(passive data types)*.
- *The technology for the first operational DAO QC system should be imported from NCEP.*

- NCEP QC contains the technology with which we are most familiar. The NCEP QC system is reasonably portable, and some of its components have been ported to different environments in the past.
- DAO has a formal agreement with NCEP which should greatly facilitate this transfer of technology.

- *A short-term requirement for the DAO QC system is that it should roughly reproduce the NCEP QC marks and corrections.*
  - This requirement allows us to introduce incremental changes with respect to the present system, and to investigate the effect of such changes.

5 Sketch of the current situation

See Figures 2 and 3 for a detailed view of the (proposed) NCEP QC system; the system is constantly in a state of flux and some of the details are not necessarily consistent with the current situation. Figure 4 is a compressed and somewhat simplified version of the current QC processing at NCEP, and indicates the data link to GEOS/DAS 2.0. The following roughly outlines the QC steps that take place prior to analysis in GEOS/DAS 2.0:

1. DAO receives observational data, supplemented with QC marks and possibly corrected, from NCEP on a daily basis.
   - The data which are currently received from NCEP by the DAO have passed through most components of the NCEP QC system, including some that utilize NCEP forecasts. As seen in Figure 4, NCEP is using an intermediate "mirror image" system at this time, with one data stream progressing through buffer storage while the other passes through files written in the so-called "Office Note 29" format (see Keyser, 1994); the second method is soon to be phased out. The principal algorithmic difference in the two branches is in the interpolation of the forecasts to the data locations. DAO accesses "Office Note 29" format files currently and so will be forced to adapt shortly when that processing path is abandoned. The link to DAO indicated in Figure 4 approximately corresponds to the box labeled BUFR 6 in Figure 3.

2. The NCEP QC'ed observational dataset is converted to an internal DAO format.

   - Conversion is done by the REPACK program, which performs certain basic sanity checks but otherwise relies on the NCEP QC marks and corrections. Based on this, observations put out by REPACK are labeled `good` or `suspect`. The current output format of REPACK is *not* ODS.

3. The observational dataset is ingested into GEOS/DAS 2.0, where final QC decisions are made on-line just prior to analysis.

   - GEOS/DAS 2.0 reads the data output by REPACK, computes observed-minus-forecast residuals, and performs a gross check on all `good` data. Data which fail the gross check are marked `suspect`. A buddy check is then performed on all `suspect` observations. Those that pass the buddy check are marked `good`. The analysis is then computed based on all `good` observations.

6 GEOS/DAS modes of operation

The GEOS Data Assimilation System will be operated in a number of modes, depending on the purpose of the assimilation. As currently planned, the modes of operation are the following:

- first-look analysis mode;
- final platform analysis mode;
- re-analysis mode.

The purposes of these modes and the main operational aspects associated with each of them are described in DAO Office Note 96-16 *(Data Assimilation Computing and Mass Storage Requirements for 1998)* and DAO Office Note 96-13 *(Requirements for DAO's On-Line Monitoring System (DOLMS) Version 1.00)*.

For QC purposes it is important to distinguish the modes of operation of the system based on the rate at which observational data are being processed. This will be roughly one month of data per day of processing in re-analysis mode, versus one day of data per day in the other three modes. An increased processing rate places higher demands on the efficiency of the on-line components of the QC system. In particular, in early versions of the GEOS/DAS QC system the components which rely on a current forecast (the *post-forecast TCA*, see section 3) may have to be simplified somewhat in re-analysis mode in order to achieve this efficiency. We will return to this issue below.

7 Implementation: GEOS/DAS 2.1

As currently planned, GEOS/DAS 2.1 is a Fortran90 serial code supporting new observation operators. It should be ready for cycled experiments by November 1, 1996.

The QC component of GEOS/DAS 2.1 will:

- be ODS-based;
- roughly reproduce the QC decisions of GEOS/DAS 2.0;
- allow passive data types [2] to flow through the system.

Figure 5 shows an information diagram for GEOS/DAS QC 2.1. Implementation requires:

[2] Data which flows through the system without being assimilated. Generally these are new data types which require evaluation before a decision can be made on how to assimilate them

1. Implementation of a generic on-line quality control module which is functionally equivalent to the GEOS/DAS 2.0 gross checks and buddy checks.

   - This requires re-engineering of the current OI routines RDHUV, DELHUV, RDMIX, DELMIX, RDSLP, and QCSLP, which
     (a) read in observational data;
     (b) compute observed-minus-forecast residuals;
     (c) perform a gross check on all data;
     (d) perform a buddy check (based on a simple interpolation) on suspect data.

     The new module will be generic in the sense that it accepts any data type. Recognition of the data type and access to observation error statistics is supported by the ODS standard.

2. Porting the major part of the NCEP QC system with the ability to utilize GEOS/DAS forecasts.

   - This includes all components shown in Figure 2 up to and including CQC.
   - The main problem that needs to be addressed in this step is the input of forecast information. All routines that require forecast information will have to be modified to be able to read a GEOS/DAS forecast from a file.

3. Modification of REPACK in order to generate output (observations and their quality marks) in ODS format.

   - This involves a minor coding change to the REPACK program. Once the port of the NCEP QC system is in place, the functionality of REPACK will be limited to converting BUFR format to ODS format.

4. Establishment and documentation of OMS formats for all data types going into the system.

   - An OMS file contains data-type dependent data attributes and must be defined for each data type; see DAO Office Note 95-01 *(Documentation of the GEOS/DAS Observation Data Stream (ODS) Version 1.0)*.

Remark on modes of operation: With the ability to utilize GEOS/DAS forecasts read from a file, various configurations of the QC system will be possible depending on the mode of operation. In modes which involve low data processing rates (such as first look analysis mode), the post-forecast components of the QC system (such as CQC) may be put on-line even if they are relatively inefficient. On the other hand, in re-analysis mode the post-forecast components may have to read stored forecast information from a file. The effect on performance—both computationally and in terms of data rejection—will have to be carefuly investigated.

Remark on radiation correction: At this time, DAO does not receive data from NCEP which has undergone radiation correction for rawinsonde data at the end of the CQC procedure, and does not have a comparable algorithm in place. In testing of the NCEP radiation correction algorithm, several problems were noted (David Lamich, private communication), and DAO may need to devise its own algorithm as a consequence. It is imperative that DAO has a radiation correction algorithm implemented in GEOS/DAS 2.1. Radiation correction is an example of a bias correction algorithm, and although it does not strictly belong to quality control it is clearly closely related to it.

8 Implementation: GEOS/DAS 3.0

As currently planned, GEOS/DAS 3.0 is a re-engineered version of GEOS/DAS 2.1 which will run efficiently in a massively parallel environment. It is supposed to be delivered by June 30, 1997.

The QC component of GEOS/DAS 3.0 will:

- roughly reproduce the QC decisions of GEOS/DAS 2.1;
- run efficiently in a massively parallel environment;
- seamlessly accomodate QC algorithms for new data types.

Implementation requires:

1. Creating a parallel version of the on-line buddy check.
   - This primarily involves creation of a search algorithm (for collecting the buddies) which performs well in a parallel environment.
2. Integration of those parts of the NCEP QC algorithms which use post-forecast data into a parallel environment. This may become a task requiring a major amount of effort.
3. Creating a generic on-line decision making algorithm.
   - This will be based on the DMA contained in NCEP's CQC. It remains to be seen whether dissection of CQC produces reusable code, or whether it makes more sense to write new code for this component.
4. Generally streamlining and re-engineering of the remaining off-line QC components.

9 Implementation: future versions

We expect significant research and development in the following areas:

- Increased sophistication of buddy checks
  - Using existing components of PSAS (solver as well as covariance models) it is not difficult to perform *optimal buddy checks* based on a local statistical analysis.

  - There are many possible strategies for buddy checks which are feasible in the context of a parallel PSAS, and which should lead to more powerful tests. By more powerful we mean an increase of the ratio between (1) the likelihood of rightly rejecting bad data and (2) the likelihood of falsely rejecting good data.

- Identification and removal of redundancies in the QC system

  - With increasing power of the buddy check, which acts as the final, fine-grained filter in the QC process, the system should be more robust to changes in the earlier stages of the process. At this point it becomes possible to systematically assess the effect of each component of the system.

- Development of off-line QC modules for new data types in collaboration with the instrument teams

  - As new data types pass through the system, the first task will be to investigate the added value of the data relative to the existing system. Only when added value has been demonstrated can the data type graduate from *passive* to *active*. Development of effective quality control (as well as bias correction) algorithms is a crucial step in this task.

Acknowledgments

A formal technical review of this document was conducted on June 6, 1996 at the Data Assimilation Office. We would like to thank Siegfried Schubert, Arlindo daSilva, and David Lamich for valuable suggestions.

References

Ballish, B. A., 1991: Quality Control Case Studies with the NMC Global Analysis System. *Preprints, Ninth Conference on Numerical Weather Prediction*, Denver, CO, American Meteorological Society, 143-146.

Collins, W. G., 1991: Complex Quality Control of Rawinsonde Heights and Temperatures at the National Meteorological Center. *Preprints, Ninth Conference on Numerical Weather Prediction*, Denver, CO, American Meteorological Society, 15-18.

Collins, W. G., and L. S. Gandin, 1996: Complex Quality Control for Observation Errors of Rawinsonde Temperatures and Heights. *Office Note xxx,* U. S. Department of Commerce, National Oceanographic and Atmospheric Administration, National Weather Service, Environmental Modeling Center.

Collins, W. G., and L. S. Gandin, 1995: Complex Quality Control of Rawinsonde Heights and Temperatures – Principles and Application at the National Meteorological Center. *Office Note 408,* U. S. Department of Commerce, National Oceanographic and Atmospheric Administration, National Weather Service, National Meteorological Center.

Collins, W. G., and L. S. Gandin, 1992: Complex Quality Control of Rawinsonde Heights and Temperatures (CQCHT) at the National Meteorological Center. *Office Note 390,* U. S. Department of Commerce, National Oceanographic and Atmospheric Administration, National Weather Service, National Meteorological Center.

Collins, W. G. and L. S. Gandin, 1990: Comprehensive hydrostatic quality control at the National Meteorological Center. *Mon. Wea. Rev.*, **18**, 2754–2767.

da Silva, A. and C. Redder, 1995: Documentation of the GEOS/DAS Observation Data Stream (ODS) Version 1.0. *DAO Office Note 95-01.* Data Assimilation Office, Goddard Space Flight Center, Greenbelt, MD 20771.

da Silva, A., K. Ekers, and A. Conaty, 1996: Requirements for DAO's On-Line Monitoring System (DOLMS) Version 1.00. *DAO Office Note 96-13.* Data Assimilation Office, Goddard Space Flight Center, Greenbelt, MD 20771.

Gandin, L. S., 1991: Two Years of Operational Comprehensive Hydrostatic Quality Control at the NMC. *Preprints, Ninth Conference on Numerical Weather Prediction,* Denver, CO, American Meteorological Society, 11-14.

Kalnay, E., M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K. C.. Mo, C. Ropelwski, J. Wang, A. Leetmaa, R. Reynolds, R. Jenne, and D. Joseph, 1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bulletin of the American Meteorological Society* **77-3**, 437 - 471.

Keyser, D., 1994: NMC Format for Observational Data (Upper-air, Single-level, Cloud Cover, Additional Reports). *Office Note 29,* U. S. Department of Commerce, National Oceanographic and Atmospheric Administration, National Weather Service, National Meteorological Center. Revised version of original 1973 edition.

Morone, L. L., 1991: Near Real-Time Monitoring of Automatic Quality Control Decisions at the National Meteorological Center. *Preprints, Ninth Conference on Numerical Weather Prediction,* Denver, CO, American Meteorological Society, 139-142.

Stobie, J, 1996: Data Assimilation Computing and Mass Storage Requirements for 1998. *DAO Office Note 96-16.* Data Assimilation Office, Goddard Space Flight Center, Greenbelt, MD 20771.

Woolen, J. S., 1991: New NMC Operational OI Quality Control. *Preprints, Ninth Conference on Numerical Weather Prediction,* Denver, CO, American Meteorological Society, 24-27.

NOTE: The NASA Technical Memoranda listed above are available by anonymous ftp from

`ftp://dao.gsfc.nasa.gov/pub/tech_memos`.

and the DAO Office Notes listed are available by anonymous ftp from

`ftp://dao.gsfc.nasa.gov/pub/office_notes`.

Modularity of the future GEOS/DAS QC

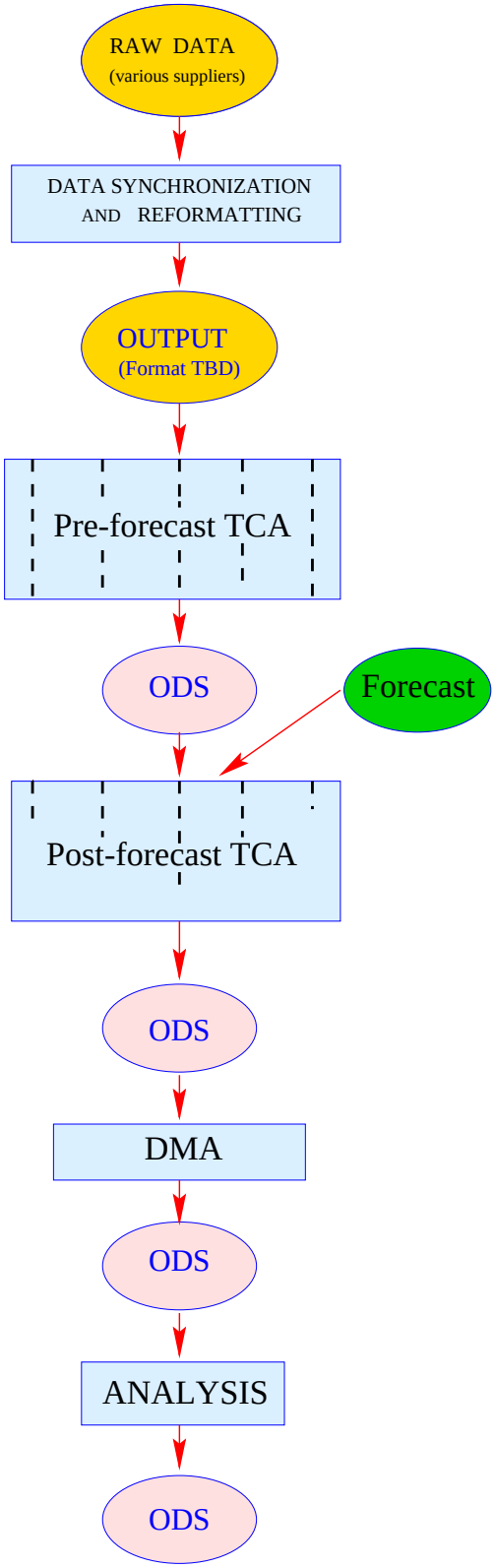

Figure 1: Modularity of the future GEOS/DAS QC

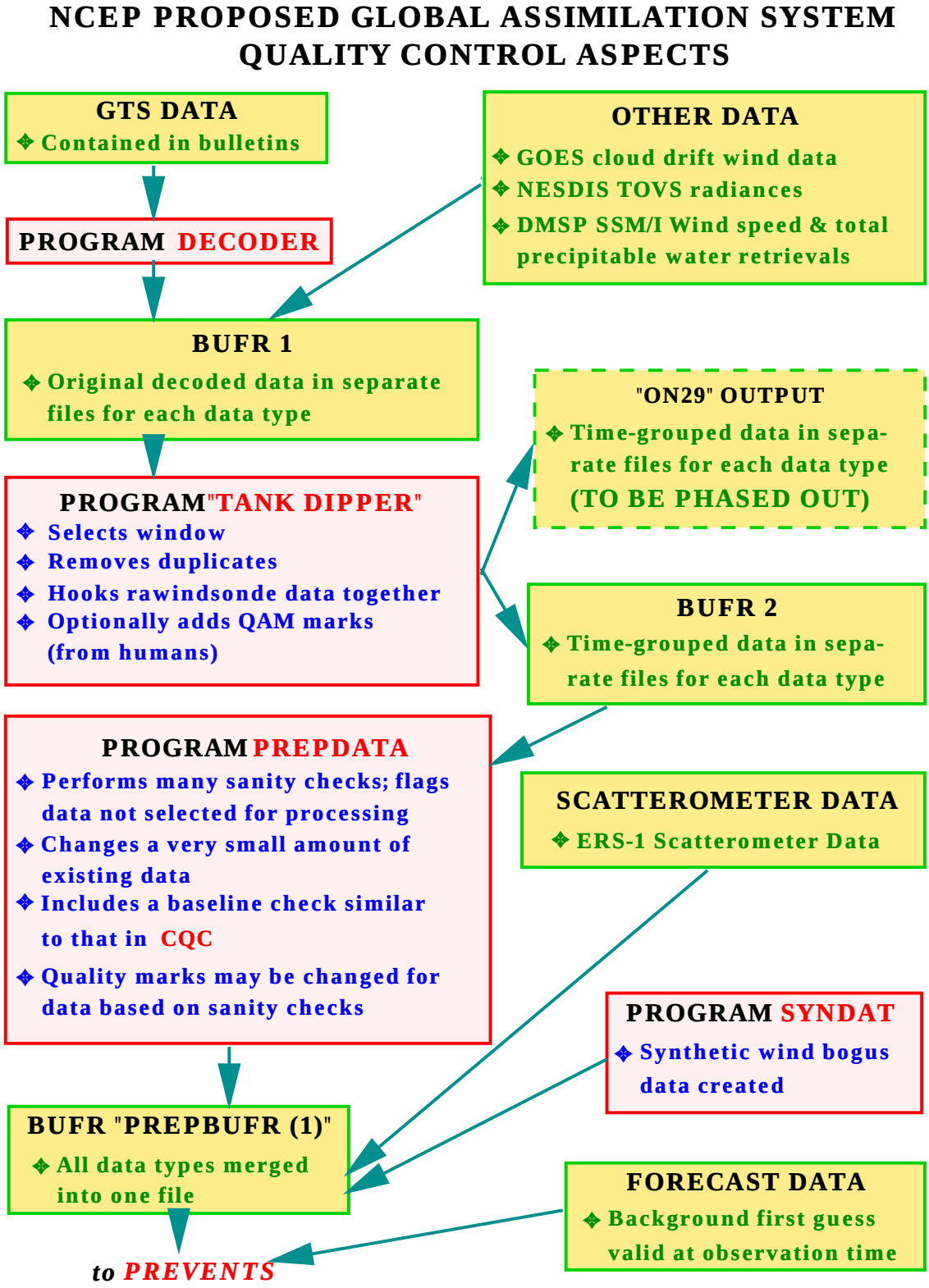

Figure 2: Detailed view of proposed NCEP QC (part 1)

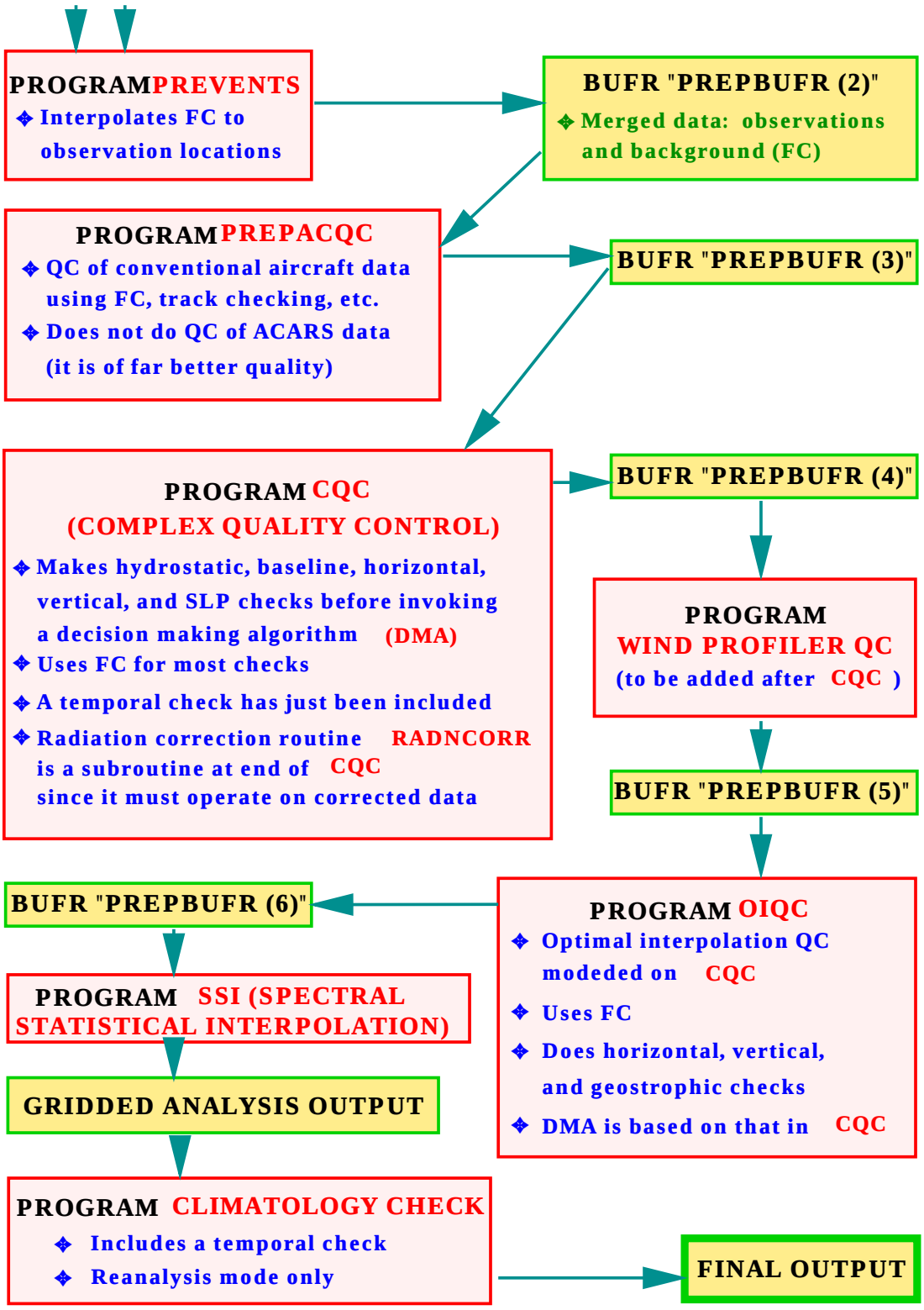

Figure 3: Detailed view of proposed NCEP QC (part 2)

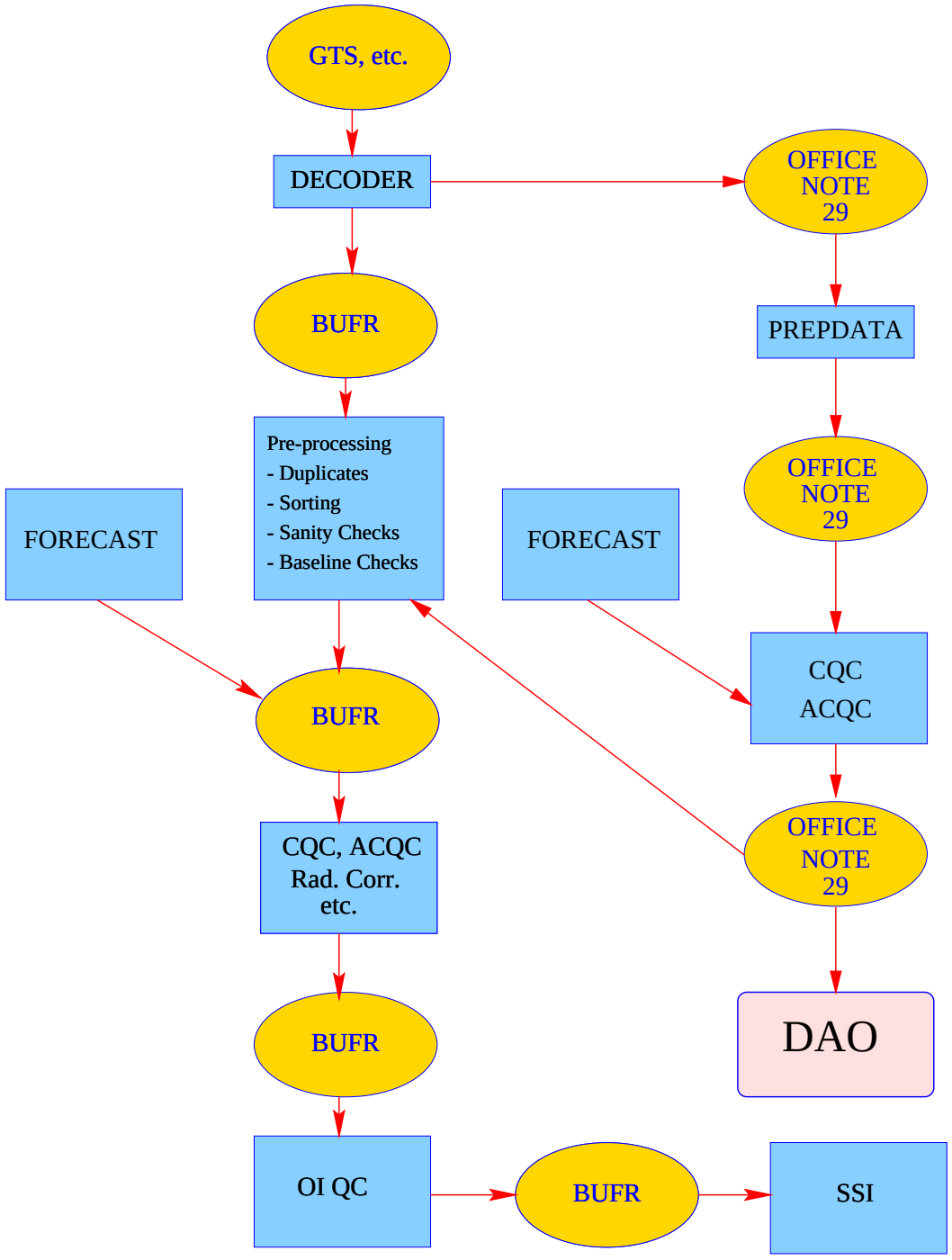

Figure 4: Simplified view of current NCEP QC and link to GEOS/DAS. NCEP uses an intermediate "mirror image" system with one data stream progressing through buffer storage while the other passes through files written in "Office Note 29" format; the second method will be phased out. The principal algorithmic difference in the two branches is in the interpolation of the forecast to the data locations.

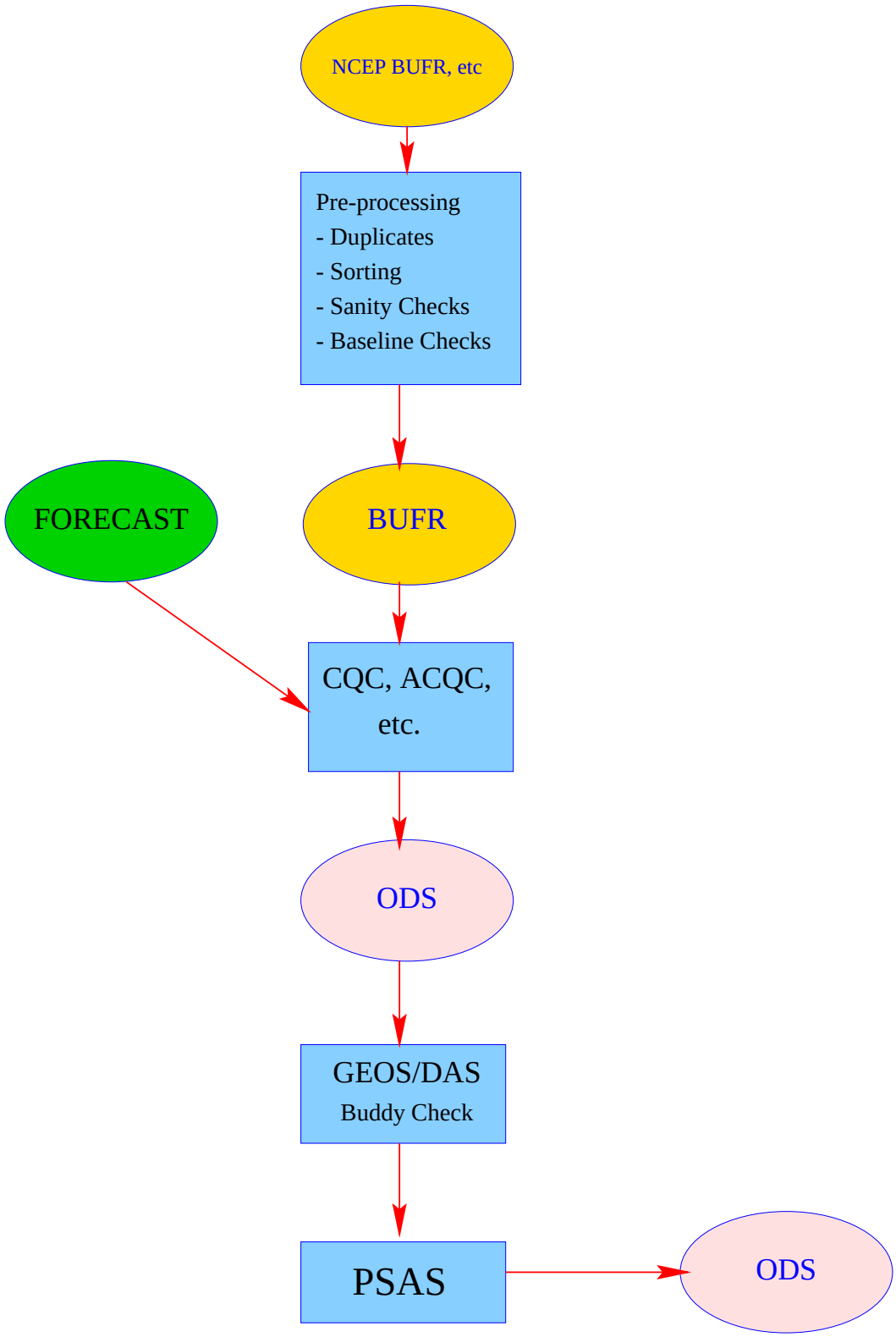

Figure 5: Schematic view of GEOS/DAS 2.1 QC